Analyzing Genomic Data with NOJAH

TAB A) GENOME WIDE ANALYSIS

NOJAH : Not Just another HeatMap   HomePage   Genome Wide Analysis   CoMMpass Data analysis   Significance of Clusters Analysis   About Us ▾

Input GW Data

Select an example dataset or upload your own with 'Load my own GW data.'

Example CoMMpass GW Expression Data ▾

⬇ Download GW Example DataSet

Step 1:

Select the example dataset or upload your own.

To view example data and format, use download button to view .csv file.

Data Subsetting

Subset GW data by:
- ☐ Variance
- ☑ Median Absoute Deviation
- ☑ Inter Quartile Range

Percentile
- ◉ Percentile Slider
- ○ Manually Enter Percentile

Percentile Value:

0   75   100

0 10 20 30 40 50 60 70 80 90 100

Step 2:

Select method of sub-setting. You can use the boxplot on the main panel to help choose the method. In the example data, MAD and IQR show relatively larger spread.

Downloads

Type the file name you would like to save subset data as :

Subset_data

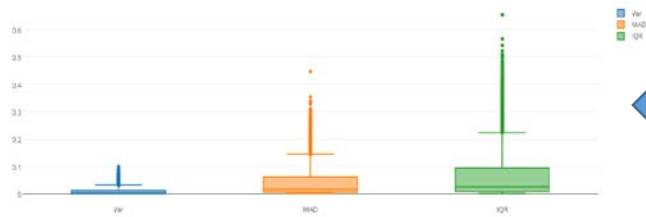⬇ Download Subset data

Save HeatMap as :

GW_subset_HM

⬇ Download HM

Step 3:

Choose a percentile cut-off to select the top most variable number of rows (genes in this case). You can also choose cut off based on the inclusion of a particular gene.

The total number of selected genes are displayed in the main panel.

Gene-wise Measures of Spread

These row-wise Variance, MAD and IQR boxplots should result if you used the example data set.

Most variable genes selected based on a percentile/ row inclusion criteria. The number of genes included displayed.
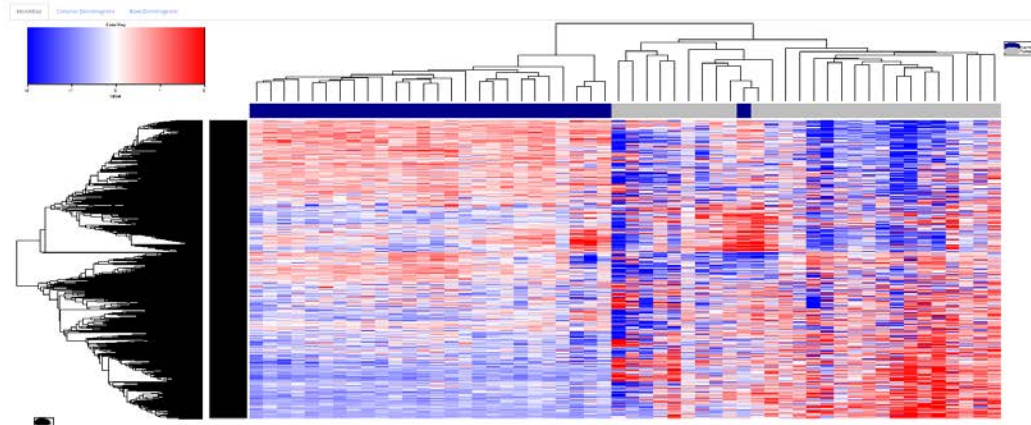
Most variable Gene Selection
To see the position of your 'gene of interest', use the 'Choose Option' drop down to the right

Choose Option:
cg00000292

Visualization of Selected Subset

A
B
C
D
E
F

Interactive HeatMap of the top most variable genes selected using the above criteria. Separate tabs display column and row dendograms. (See tutorial for part C for detailed run-through of heatmap options)

A. Data is Z-scored before input into the heatmap.2 function.
B. Data can be normalized by row, column or both
C. Scale is set from -2 to 2 but can be changed by the user
D. Choose clustering and distance measure
E. Supervised row-wise or column wise clustering can be selected using FALSE option. Row and column labels can be displayed using TRUE option
F. Change color of HeatMap using the high, mid and low colors

TAB B) CoMMpass DATA ANALYSIS

Note: The same steps apply to Expression, Variant and Copy Number tabs.
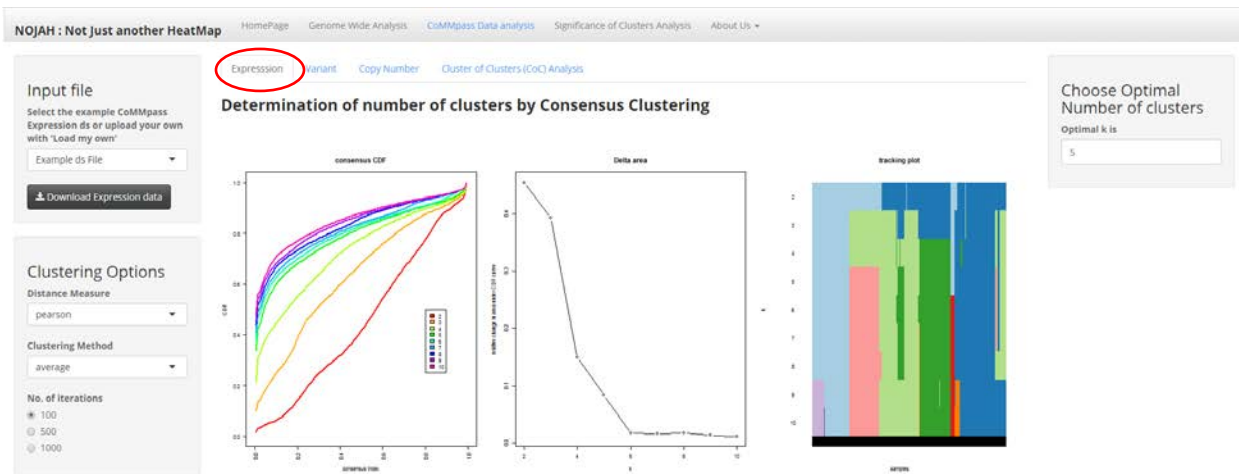
Step 1:
Select a pre-filtered RNASeq Expression CoMMpass IA9 data file or input your own coMMpass expression file.

Step 2:
Choose distance and clustering measures to perform **consensus clustering** using the 'ConsensusClusterPlus' Bioconductor package, to predict optimal number of Sample clusters for the data.

Step 3:
Choose the no. of iterations. Default is set to 100 for faster computing. In practice, set this to 1000 iterations.
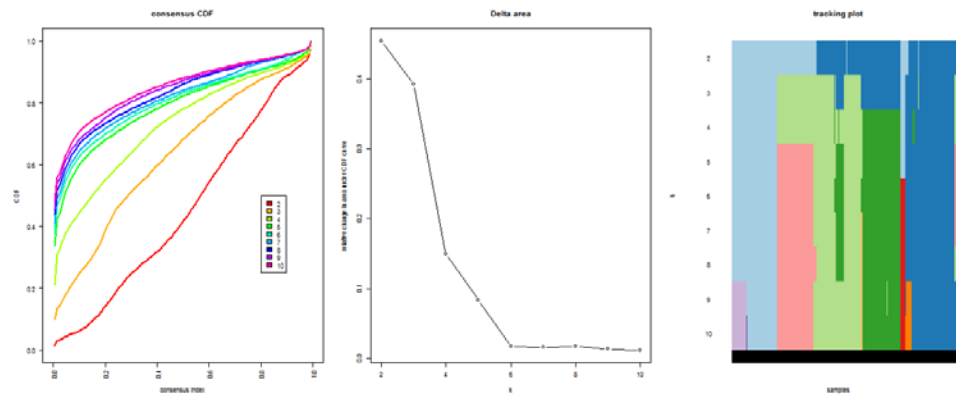
Static parameter settings used:
- item resampling = 80%
- gene resampling = 80 %
- maximum evaluated k = 9
- clustering algorithm = Agglomerative Hierarchical clustering algorithm "hc"
- Same clustering method is applied to both inner Linkage and final Linkage parameters.

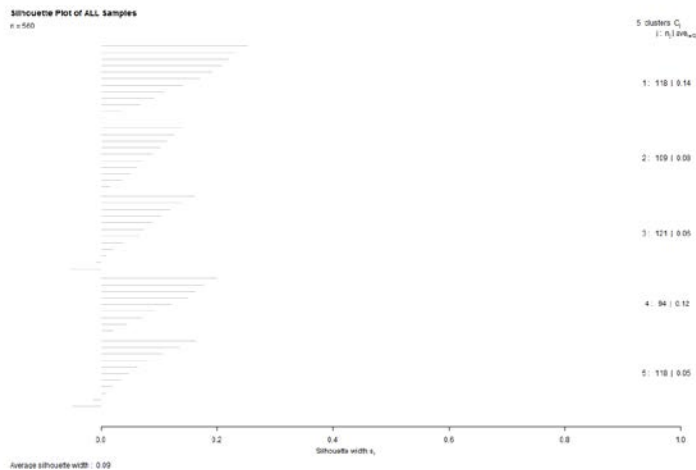## Determination of number of clusters by Consensus Clustering

Silhouette Plot

These plots will be displayed using the parameter setting above.

Consensus CDF, Delta area plot are from the output results of the 'ConsensusClusterPlus' package. Along with the consensus matrix heatmap (available for download), they will help the user determine the optimal number of clusters in the data.

For this Expression data, five optimal clusters are predicted. The user can change the optimal clusters using the right panel.

Silhouette Plot can further help confirm the identification of the number of clusters visually. The larger the average silhouette width, the more reliable the cluster structures
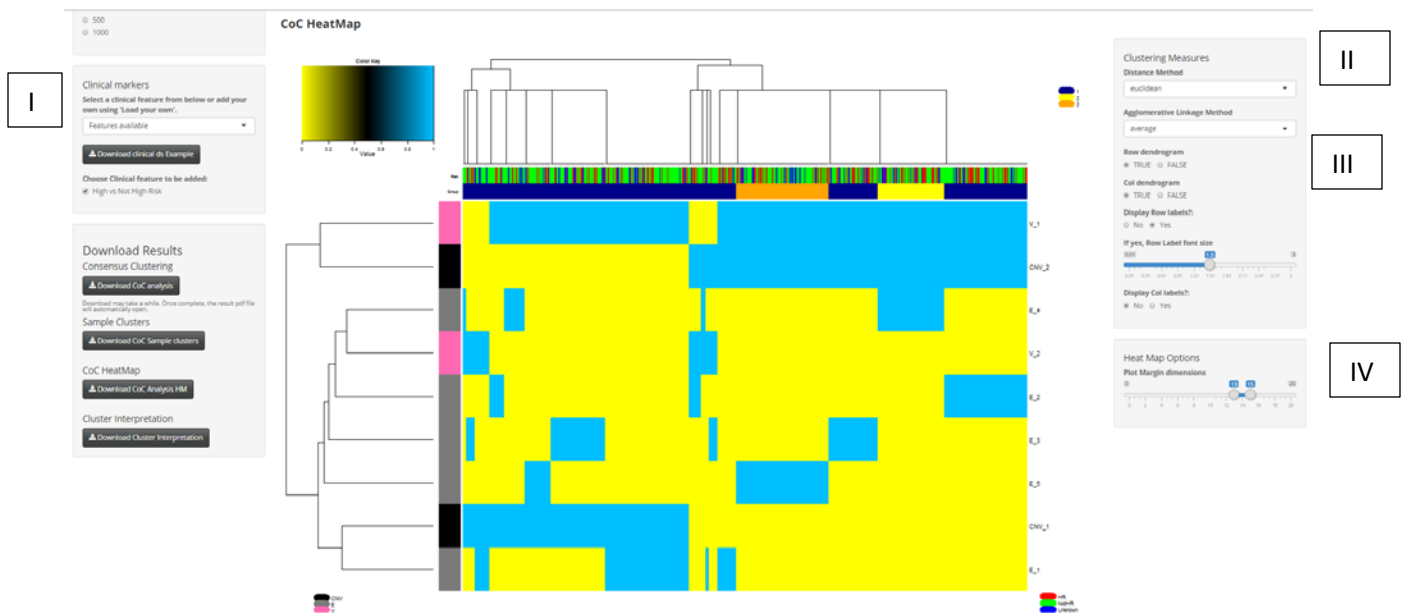
CLUSTER OF CLUSTER ANALYSIS
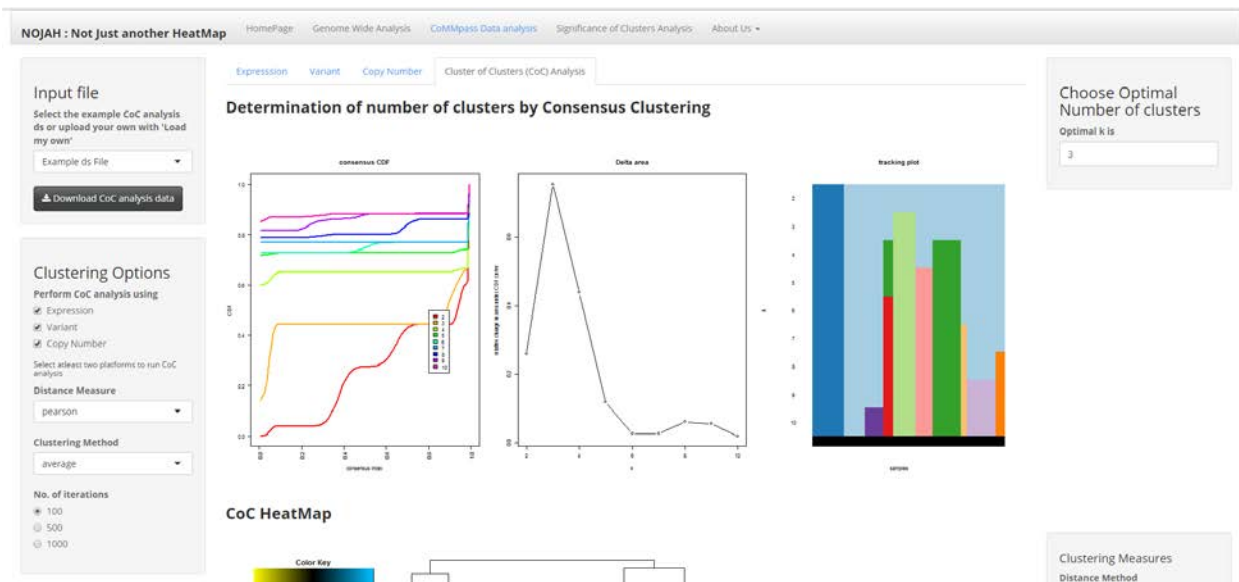
Step 1:
Select the computed number of optimal RNASeq Expression, Variant and CNV data from the previous 3 tabs or upload the your own clusters using the load my own option. The user-uploaded cluster data should be in the same format as the example data. Example data is available for download. In addition, the same patients should be used for each platform in order to be used for CoCA.

Step 2:
Select the platforms to base CoC analysis. User should select at least two platforms.

Select other parameter setting in the similar fashion to the previous tabs to run 'ConsensusClusterPlus' package.

Interactive HeatMap for Cluster of Cluster Analysis.

1-0 Transformed matrix data based on the individual platform clusters is used as input into the modified heatmap.2 function.

I.   Add Single or multiple clinical feature(s) as bars just below the dendrogram. As an example, sample risk status is displayed above the predicted consensus cluster bar and can be downloaded using the download button.

II.   Choose clustering and distance measure

III.   Supervised row-wise or column wise clustering can be selected using FALSE option. Display Row and column labels using TRUE option. Adjust size of the labels using the slider.

IV.   Adjust Plot margins using the slider.

Cluster Interpretation
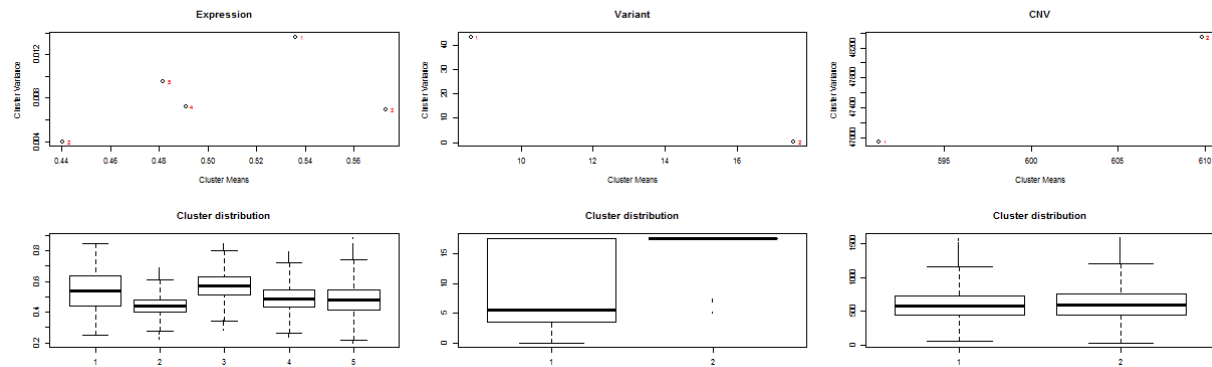
Interpretation of CoC Analysis Cluster HM based on the individual platform clusters.

Variance vs the mean plot of the lower triangular distance matrix serves as a relative measure of each cluster relative of the others within the same platform.

Boxplot of the individual clusters also helps determine which cluster has a relatively higher or lower median Expression (or median proportion or median CNV segment mean). The spread among the clusters is also informative.

TAB C) SIGNIFICANCE OF CLUSTER ANALYSIS
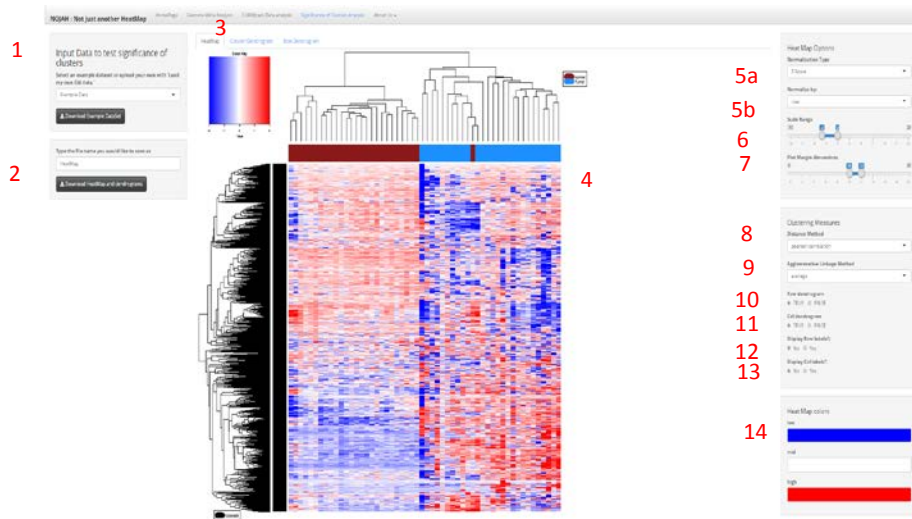
1: Select dataset of interest. Using the dropdown, you can choose the example or upload your own. If uploading your own, format data in same format as in the example file.

2: Download example data using download button to view contents/formatting of example file.

**3:** If example file is chosen, Heatmap automatically displayed in the HeatMap tab.

4: HeatMap created using Z-score 'row' normalization, 'Pearson correlation' distance and 'average' agglomerative linkage method (i.e. default settings). Depending on dataset may take several minutes to load.

5a, b: Select a different normalization method you'd like for the data using drop down options. Each time a different type is chosen, the heatmap will be updated.
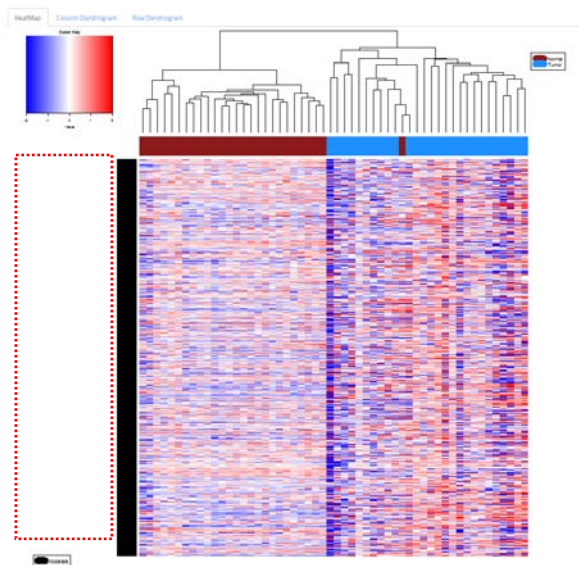

6 (optional): Drag slider to change scale range for the colors. Heatmap will be updated on movement.

7: Select the Plot margins. If column dendrogram overlaps the legend, increase both margin points and vice – versa until desired.
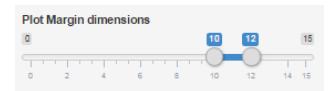

8, 9: Select Distance method and linkage method of choice using the drop down options. Each selection will display modified heatmap.

10, 11: Select either to display Row dendrogram or not. If FALSE is chosen, row dendrogram will disappear and data will not be ordered based on means. Same applies to Column dendrogram.
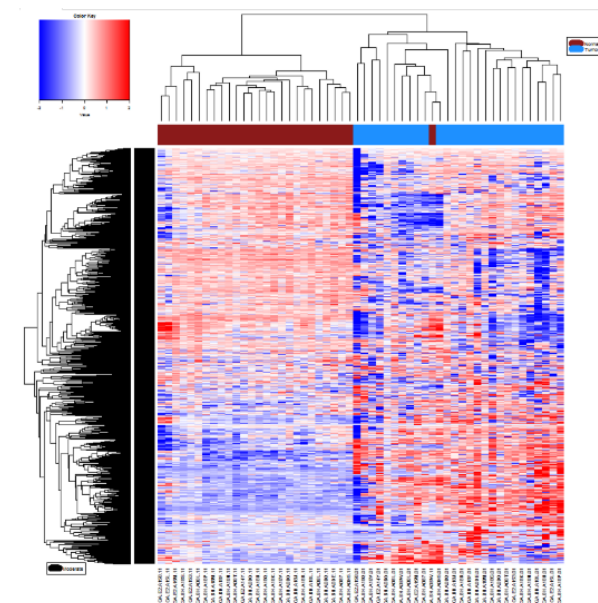

12, 13: Select Display Row labels = 'Yes' to see the corresponding CpG sites. Additional slider appears to select, font size. Same applies to Sample labels.
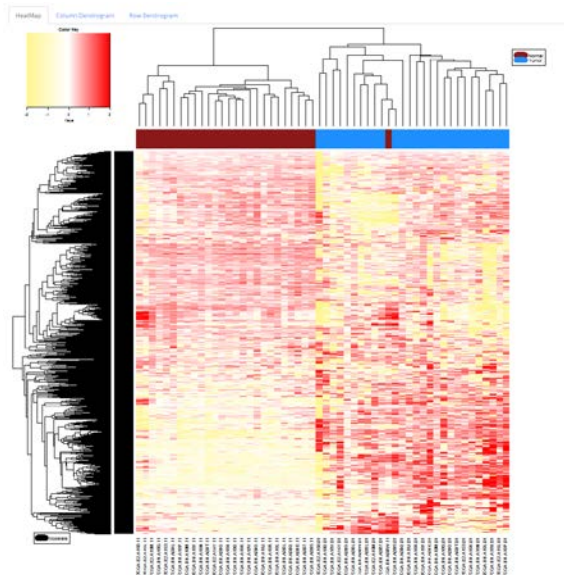
14: Select color scheme. Red-Black-Green is typically used for Expression data and Blue-White-Red is used to represent methylation data. Heatmap will update as soon as color is chosen. After choosing desired color(s), click anywhere on screen to come out of color selection panel.

15: Input file name and click on Download button to save heatmap and the corresponding row and column dendrograms in pdf format as shown below using Chrome browser.
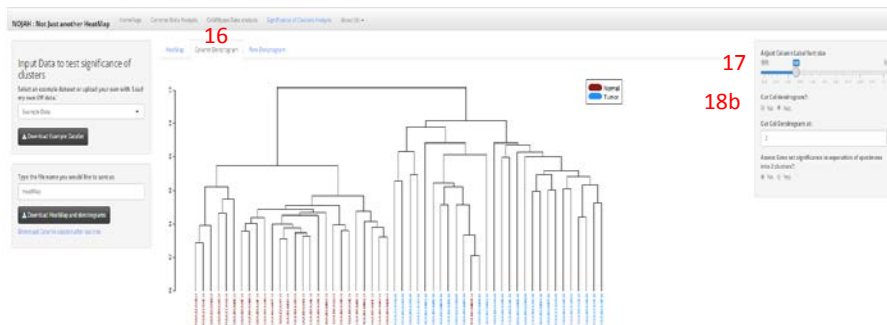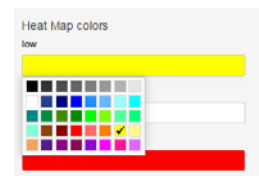
16: View in column dendrogram tab

17: Slider to adjust font size of the column dendrogram labels

18 a, b: a. Option to cut the tree. b. If yes is chosen, user is asked at which position they want to cut the tree (default at 2)

When selected, a table will appear that classifies Samples, their Groups, and their corresponding clusters.

Use the drop down on upper left [ ], to display 5/10/All rows of the table.

19: Option to assess gene set significance in separation of the two clusters (Tumor vs Normal). Applicable only when >=2 clusters are available for analysis.

When 'Yes' is selected, parameters for Monte Carlo p-value estimation will be made available.

18a Please select Cut Col dendrogram?: = 'Yes' to display column clusters. Also select value at which you would like to cut the col dendrogram (default is at k= 2)

19 Would you want to assess gene set significance in the separation of specimens into two clusters? (Yes/No)

19a: Select Sampling dataset for bootstrap. An example Methylation Sampling data is available or user can input their own (up to 75 MB is allowed). Large .csv and .txt files can be converted to .RDS file contain file size within 75 MB limit.

19b: Choose Sample size of the data for bootstrap. Use a size that does not exceed the original sampling data itself.

19c: Select number of iterations you wish to perform. A good practice is to perform at least 1000 iterations for accuracy of analysis.

19d: Once all options are selected, press 'Go' button to start analysis.

After approximately 10 seconds, a progress indicator will appear to track the time remaining for the analysis to be completed.
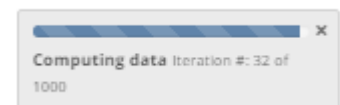
p-value results from the boot strap approach for calculation significance of clusters using Fisher's exact test will be displayed under the table along with the interpretation.
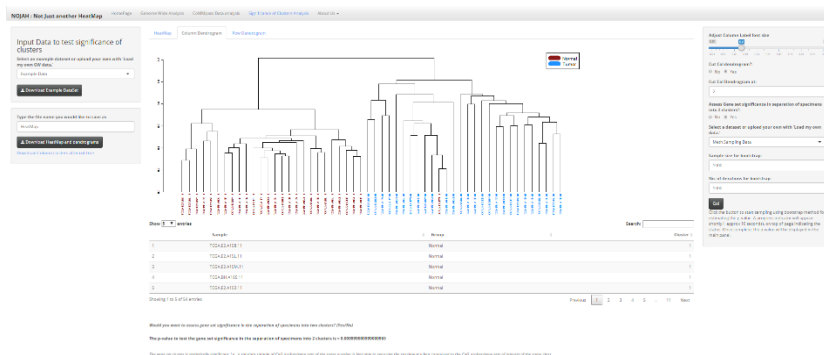
20: To download the p-value results as well, input the file name and click on Download button. The heatmap and the corresponding row and column dendrograms followed by the p-value results will be downloaded in pdf format.

To download the table for the classification of samples by clusters, click on link and the table will be saved as a .csv file.

Similar analysis can be performed on Row Dendrogram, provided you have at-least two row groups.